

OCR A Level

Computer
Science

H446 – Paper 1



Search Engine Indexing and PageRank

Unit 5

Networks and web
technologies



PG ONLINE

Objectives

- To understand how web pages are indexed by search engines
- To understand the PageRank algorithm
- To be able to interpret and apply the PageRank algorithm to a given scenario

Larry Page and Sergey Brin

- Founders of Google

- What impact did their creation have on users of the World Wide Web?



Google Search

I'm Feeling Lucky



Search engines

- It is impossible to determine how many web pages exist on the World Wide Web
 - It is estimated that the number is in the trillions
 - How can we find the information we need?



Search engines

- Search engines are systems that locate resources (web pages, files, pictures) on the World Wide Web
- Search engines make it easier to find the relevant resources in an efficient way

- Examples include:

- Google
- Bing
- Yahoo
- Baidu



Search engine indexing

- Search engines keep a record of the resources located on the World Wide Web
 - This is known as an **index**
 - The process of creating an index includes using a piece of software called a **web crawler** or **spider**



Web Crawlers

- **Web crawlers** are Internet bots that continuously crawl the web to discover and record publicly available web pages
 - Web crawlers look at web pages and follow the hyperlinks located on those pages
 - The web crawler then continues to follow the hyperlinks on the proceeding pages
 - It does this for billions of web pages. The web crawler keeps a record known as the **index**



Web Crawlers

- The information stored by the web crawler includes:
 - The URL of the resource
 - The content of the resource (ie. Text on a web page)
 - The last time the resource was updated
 - The quality of the resource (ie. How credible it is)
- The information will be stored on the search engine's database and form part of the index
- When you perform a web search via a search engine, you are actually searching the index – not the World Wide Web



Index

- The search engine index contains entries for trillions of web pages
 - The size of Google's index is over 100 petabytes (100,000 Terabytes or 100,000,000 Gigabytes) in size
 - Think of the World Wide Web as a book with a over a trillion pages. The index keeps track of what each page of the book contains



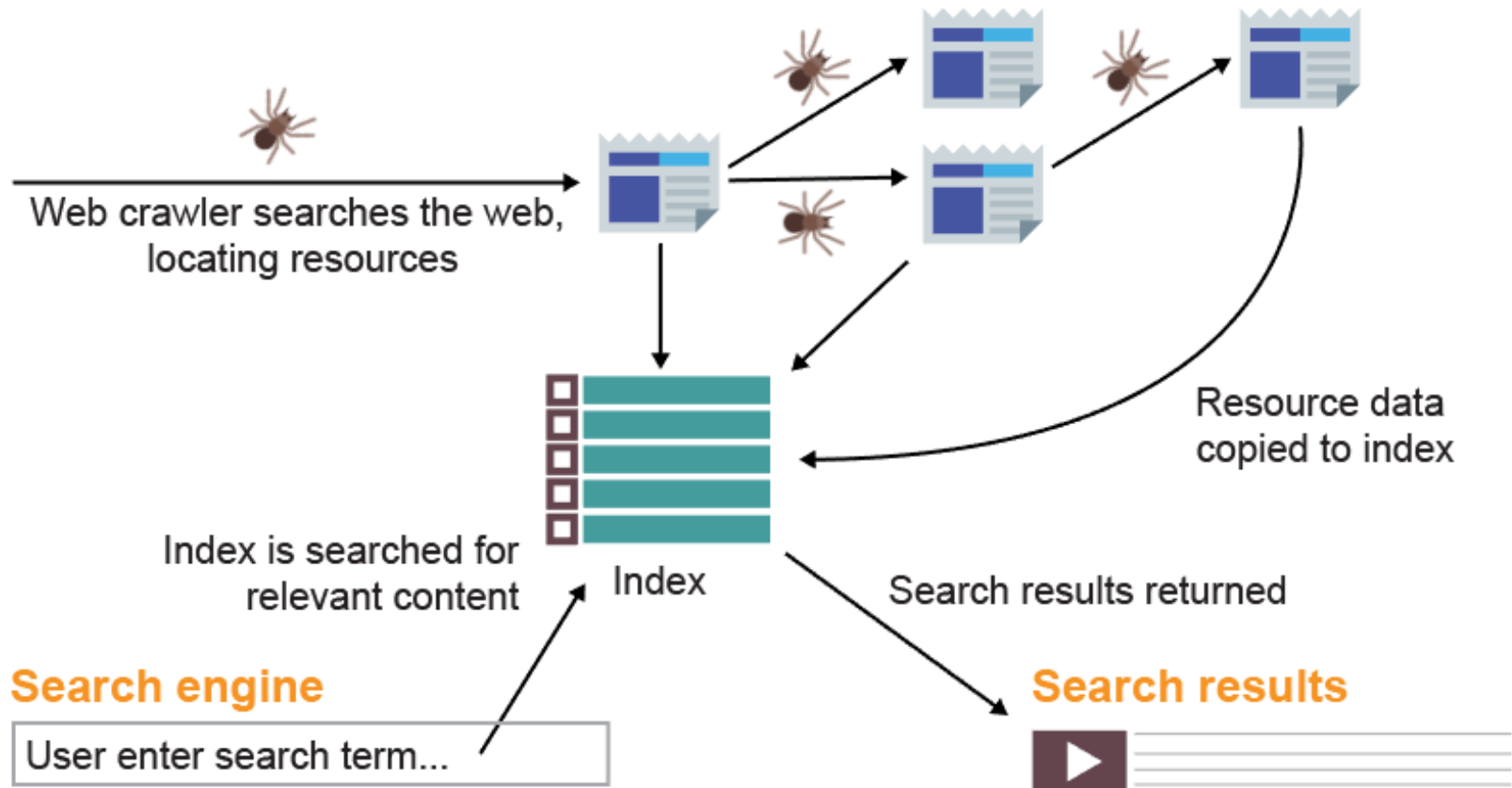
Meta tags

- Meta tags describe the content of a web page
 - Web developers can place **meta tags** inside HTML pages to make the page more likely to be found
 - Hidden from users but discoverable by web crawlers

```
1 <!doctype html>
2 <html>
3 <head>
4 <meta http-equiv="Content-Type" content="text/html;
  charset=utf-8">
5 <TITLE>Tolpuddle Martyrs</TITLE>
6 <META NAME="Keywords" CONTENT="martyr, tolpuddle, farm,
  worker, labourer, dorset, loveless, 1834, union, liberty,
  australia">
7 <META NAME="Description" CONTENT="In the 1830s life in rural
  villages like Tolpuddle was hard and getting worse. Farm
  workers could not bear yet more cuts to their pay. Some fought
  back against land owners and formed the first trade unions.">
8 </head>
9 <body>
10 </body>
11 </html>
```



Searching the web



Activity

- Open a web browser and go to www.google.com
- Enter the search term **‘Are polar bears losing weight?’**
 - How many results did the search engine return?
 - How quickly did the search engine return the results?
 - How are the results listed?



Index

- When we enter a search term, the search engine looks through the index and identifies every resource that contains those search terms
- After searching the index, the search engine will present the results to the user
 - How does the search engine calculate which web pages are relevant?
 - How does the search engine know how to list the pages?



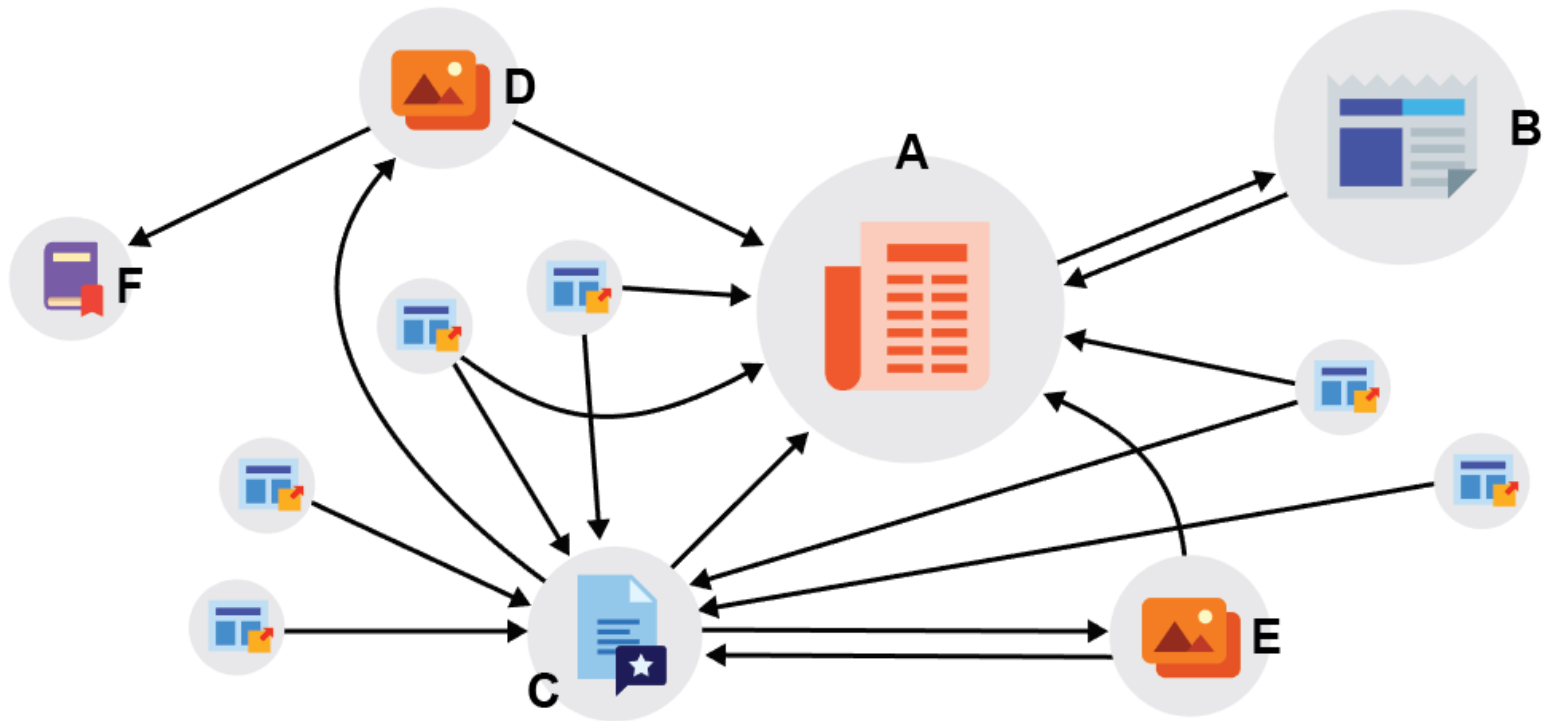
PageRank

- PageRank was developed to list search results in the order and rank of usefulness and relevance
- Created by the founders of Google, Sergey Brin and Larry Page
 - Until PageRank, search results were generally ranked in the order of how many times the search item appeared on the web page
 - The algorithm takes into account inbound links to a website to determine how useful a web page is



PageRank

- Using PageRank, **B** has a higher page rank than **C** because it is a more authoritative



PageRank Algorithm

- The original PageRank algorithm is:

$$\mathbf{PR(A) = (1-d) + d (PR(Ti)/C(Ti) + \dots + PR(Tn)/C(Tn))}$$

- **PR(A)** is the PageRank of page A
- **PR(Ti)** is the PageRank of pages Ti which link to page A
- **d** is the damping factor
- **C(Ti)** is the number of outbound links on page Ti

PageRank Algorithm

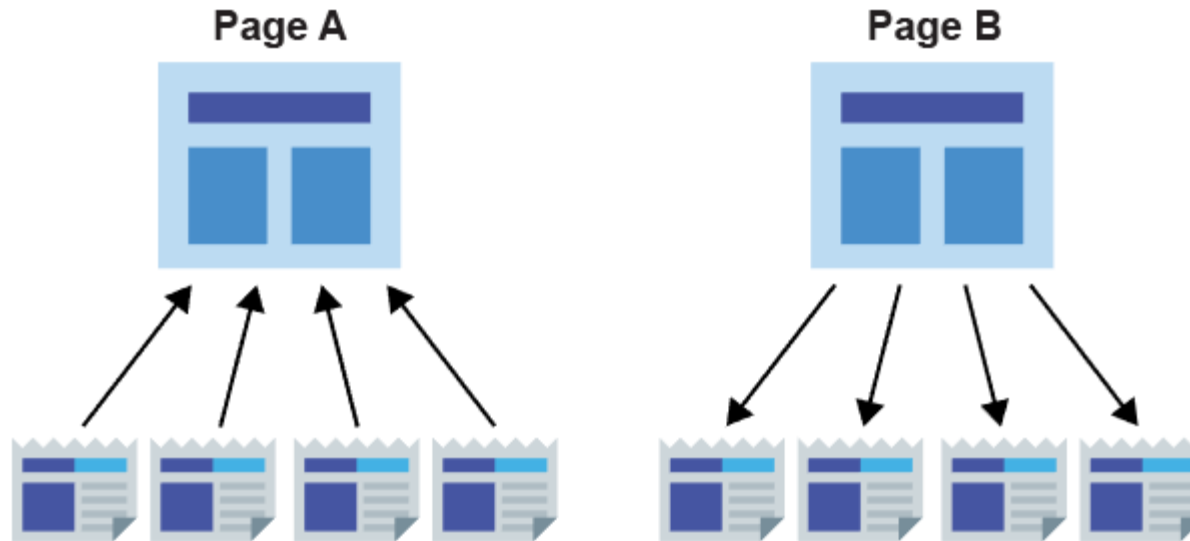
$$PR(A) = (1-d) + d (PR(Ti)/C(Ti) + ... + PR(Tn)/C(Tn))$$

- The algorithm does not rank websites as a whole
- Each web page has its own PageRank
- The PageRank of page A is defined by the PageRanks of those pages linked to page A
- The damping factor (**d**) is the probability of a random web browser reaching a page. This value is usually set to 0.85



PageRank

- The importance of a web page is determined by the number of inbound links from other pages

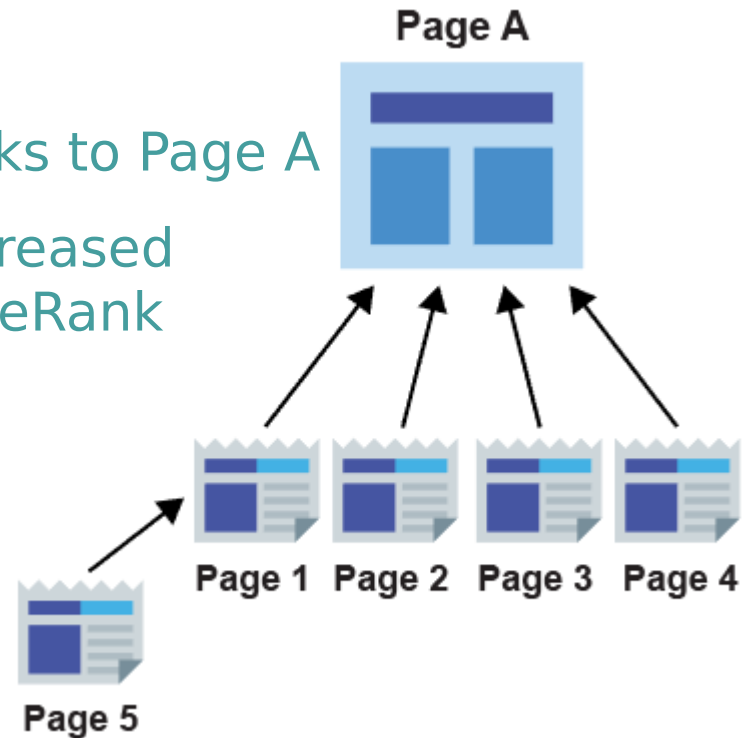


- In this scenario, **Website A** would have a higher PageRank



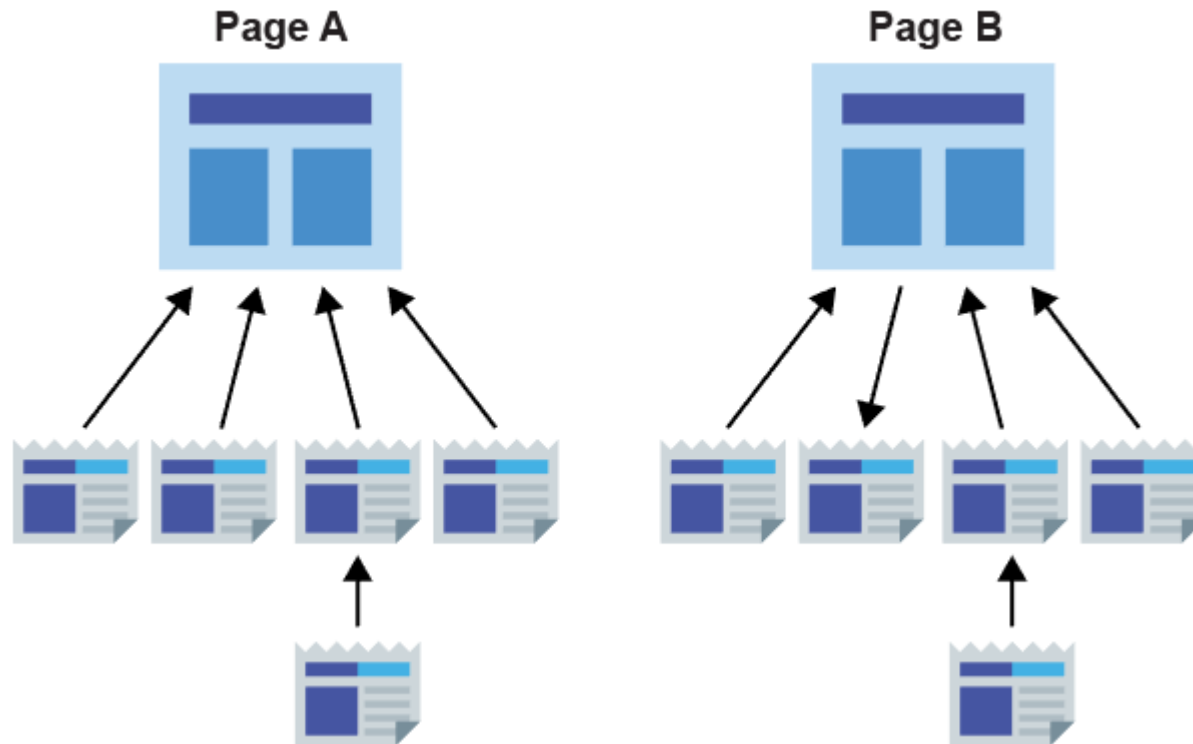
PageRank

- The quality of the links also affect the PageRank
 - In this scenario, Page 5 links to Page 1, which links to Page A
 - Page A's PageRank is increased due to the increased PageRank of Page 1



Assessing PageRank

- Which website might have the higher PageRank?



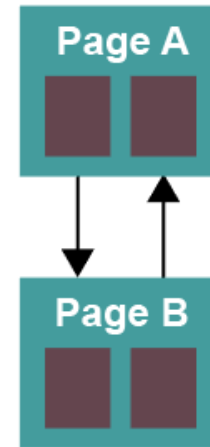
PageRank factors

- There are around two hundred factors that affect a PageRank
- These include:
 - **Domain name** – relevance to the search item
 - **Frequency of search term** in web page
 - **Age** of web page
 - **Frequency of page updates**
 - **Magnitude** of content updates
 - **Keywords** in <H1> tags



PageRank Example

- Two web pages – A and B
 - Page A links to Page B
 - Page B links to Page A
 - Since we don't know the PageRank of either page, we'll assume that it is 1



$$\text{PR}(A) = (1 - 0.85) + 0.85 \left(\frac{1}{1} \right)$$

= 1.00

PageRank
of page A

Damping
factor (d)

Damping
factor (d)

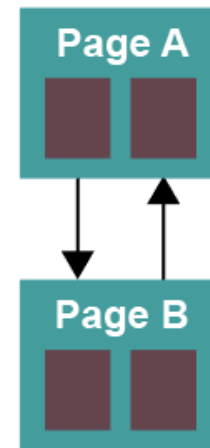
Initial
guess of
Page B's
PageRank

Number of
outbound
links on
Page B



PageRank Example

- We have calculated an initial PageRank value for Page A (**1.00**), so we can use this in the calculation for Page B's PageRank



$$PR(A) = (1 - 0.85) + 0.85 (1 / 1) =$$

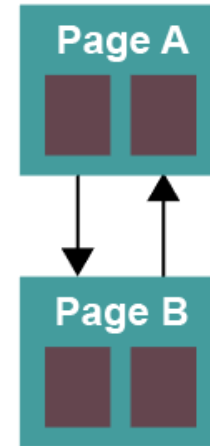
1.00

$$PR(B) = (1 - 0.85) + 0.85 (\mathbf{1.00}/1) =$$

1.00

PageRank Example

- We can now go back to our calculation for Page A, as we have a more accurate value for Page B's PageRank (**1.00**)

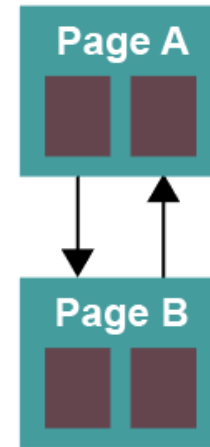


$$PR(A) = (1 - 0.85) + 0.85 (\mathbf{1.00} / 1) = \mathbf{1.00}$$

$$PR(B) = (1 - 0.85) + 0.85 (\mathbf{1.00} / 1) = \mathbf{1.00}$$

PageRank Example

- The sum of each PageRank calculation will eventually add up to the number of web pages in that particular scenario
 - In this case, $1.00 + 1.00 = 2$ pages



$$PR(A) = (1 - 0.85) + 0.85 (1.00 / 1) = 1.00$$

$$PR(B) = (1 - 0.85) + 0.85 (1.00 / 1) = 1.00$$

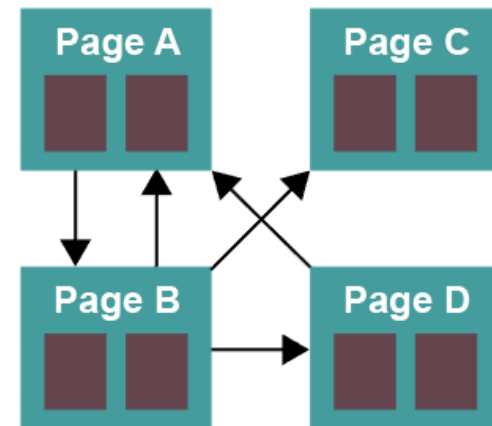
PageRank

- The PageRank value is accurately calculated after several iterations of running the formula
- Each iteration produces a more accurate PageRank value
- The higher the value is, the more relevant the page, in theory, and therefore the higher the page will appear on the search results



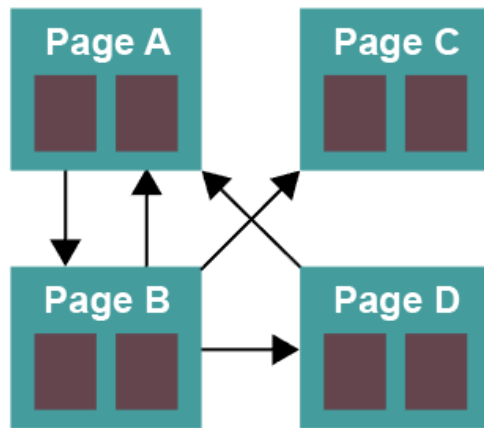
PageRank Example 2

- Four web pages – A, B, C and D
 - Page A links to Page B
 - Page B links to Page A, C and D
 - Page C does not link to any other page
 - Page D links to Page A



PageRank Example 2

- We begin with our assumption that the PageRank for each page is 1



Factor in Page D, as it provides an incoming link to Page A

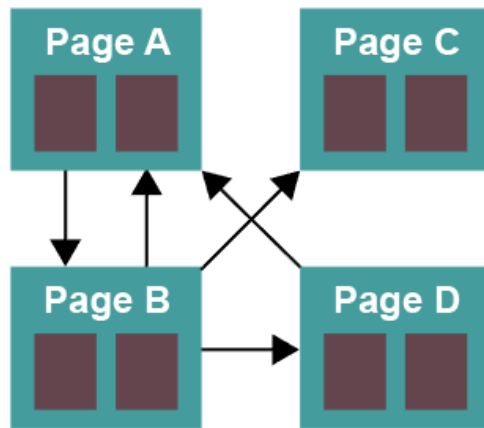
$$\text{PR}(A) = (1 - 0.85) + 0.85 \left(\frac{1}{3} \right) + \left(\frac{1}{1} \right) =$$

1.43


Page B has 3
outbound links

PageRank Example 2

- We now have an initial PageRank for Page A
- Use this in the formula for Page B:



Page A has
1 outbound
link

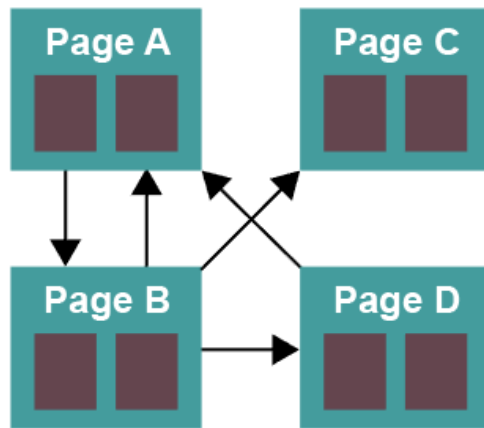


$$PR(B) = (1 - 0.85) + 0.85 (1.43 / 1) =$$

1.37

PageRank Example 2

- Page B's PageRank is used in the calculation for Page C since it has an inbound link



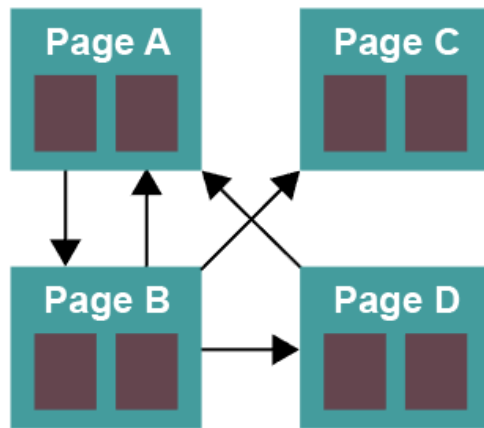
Page B has 3
outbound
links

$$PR(C) = (1 - 0.85) + 0.85 \left(\frac{1.37}{3} \right) =$$

0.54

PageRank Example 2

- Page C is not used in Page D's calculation as these two pages do not link in any way



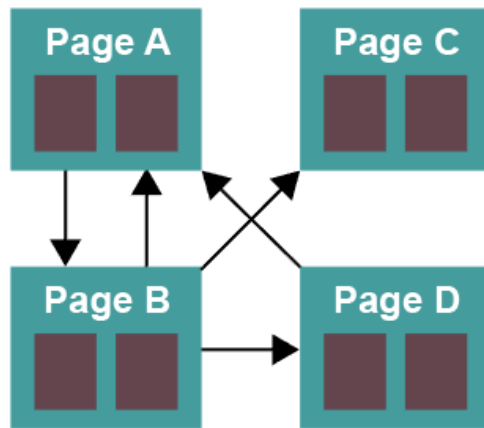
Page B has 3
outbound
links

$$PR(D) = (1 - 0.85) + 0.85 (1.37 / 3) =$$

0.54

PageRank Example 2

- After a single iteration we have the following PageRank for each web page:

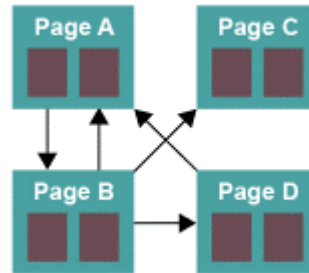


- Page A: 1.43
- Page B: 1.37
- Page C: 0.54
- Page D: 0.54

- You can see that Page A and Page B are initially more 'relevant' than C and D as they have higher values

PageRank Example 2

- We can iterate the formula with updated values for the PageRank – we no longer have to guess



$$\text{PR}(\text{A}) = (1 - 0.85) + 0.85 (1.37 / 3) + (0.54/1) = \mathbf{1.07}$$

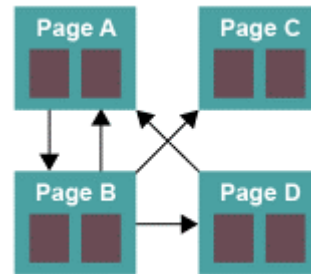
$$\text{PR}(\text{B}) = (1 - 0.85) + 0.85 (1.07 / 1) = \mathbf{1.06}$$

$$\text{PR}(\text{C}) = (1 - 0.85) + 0.85 (1.06 / 3) = \mathbf{0.45}$$

$$\text{PR}(\text{D}) = (1 - 0.85) + 0.85 (1.06 / 3) = \mathbf{0.45}$$

PageRank Example 2

- Another iteration of the formula gives us:



$$\text{PR}(\text{A}) = (1 - 0.85) + 0.85 (1.06 / 3) + (0.45/1) \\ = \mathbf{0.90}$$

$$\text{PR}(\text{B}) = (1 - 0.85) + 0.85 (0.90 / 1) = \mathbf{0.92}$$

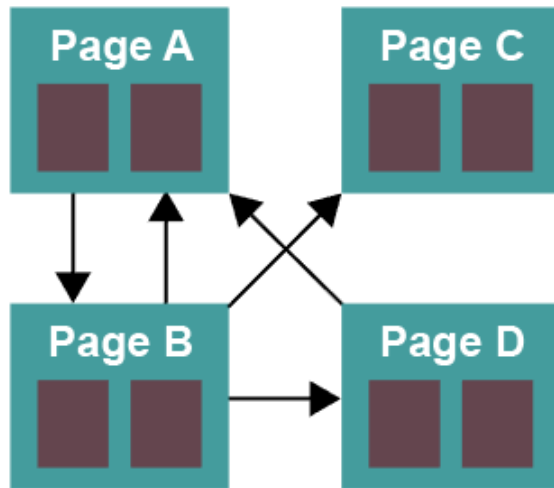
$$\text{PR}(\text{C}) = (1 - 0.85) + 0.85 (0.92 / 3) = \mathbf{0.41}$$

$$\text{PR}(\text{D}) = (1 - 0.85) + 0.85 (0.92 / 3) = \mathbf{0.41}$$



PageRank Example 2

- After forty iterations of the formula, we are presented with the following PageRank for each web page



PR(A) =

0.64

PR(B) =

0.69

PR(C) =

0.34

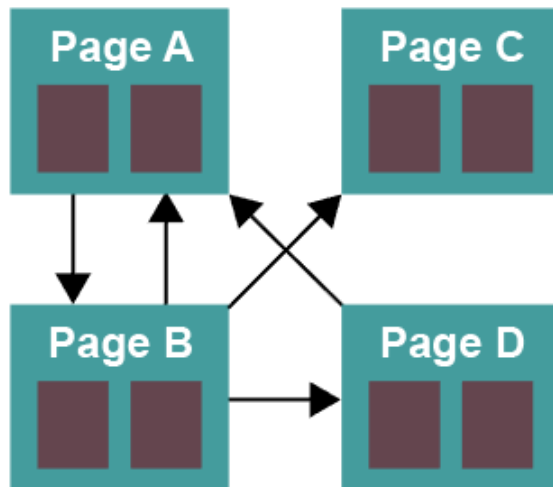
PR(D) =

0.34



PageRank Example 2

- As we can see, Page B has the highest PageRank, and therefore will be at the top of the search results in this particular instance
- Page A will be listed second, with Page C and D joint third



$$\begin{aligned} \text{PR}(\text{A}) &= 0.64 \\ \text{PR}(\text{B}) &= 0.69 \\ \text{PR}(\text{C}) &= 0.34 \\ \text{PR}(\text{D}) &= 0.34 \end{aligned}$$

PageRank Summary

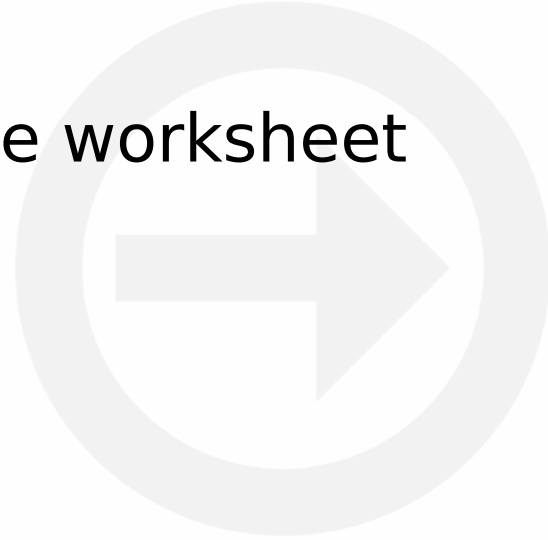
- We have to guess the initial PageRank value – we can use any number, but 1 is a good value to start with
 - Check how many incoming links there are to the page you are applying the formula to
 - Substitute the 'guess' value after the first calculation in future iterations

$$\mathbf{PR(A) = (1-d) + d (PR(T_i)/C(T_i) + \dots + PR(T_n)/C(T_n))}$$



Activity

- Complete **Activity 1** on the worksheet



Copyright

© 2016 PG Online Limited

The contents of this unit are protected by copyright.

This unit and all the worksheets, PowerPoint presentations, teaching guides and other associated files distributed with it are supplied to you by PG Online Limited under licence and may be used and copied by you only in accordance with the terms of the licence. Except as expressly permitted by the licence, no part of the materials distributed with this unit may be used, reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic or otherwise, without the prior written permission of PG Online Limited.

Licence agreement

This is a legal agreement between you, the end user, and PG Online Limited. This unit and all the worksheets, PowerPoint presentations, teaching guides and other associated files distributed with it is licensed, not sold, to you by PG Online Limited for use under the terms of the licence.

The materials distributed with this unit may be freely copied and used by members of a single institution on a single site only. You are not permitted to share in any way any of the materials or part of the materials with any third party, including users on another site or individuals who are members of a separate institution. You acknowledge that the materials must remain with you, the licencing institution, and no part of the materials may be transferred to another institution. You also agree not to procure, authorise, encourage, facilitate or enable any third party to reproduce these materials in whole or in part without the prior permission of PG Online Limited.